DOCUMENT RESUME

ED 222 525 TM 820 672

AUTHOR Jones, Douglas H.
TITLE Foundations for the Mathematical Notion of Information in Item Response Theory and Robust Ability Estimation.
PUB DATE [81]
NOTE 10p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Ability Identification; Estimation (Mathematics); *Latent Trait Theory; Mathematical Models; Maximum Likelihood Statistics; Probability; *Statistical Analysis
IDENTIFIERS *Information Function (Tests); Robustness

ABSTRACT

A mathematical setting based on a statistical sampling probability mechanism is described. In this setting, a mathematical meaning is given for the information function; and it becomes possible to study the relative merits of various ability-estimating procedures. The maximum likelihood estimation procedure under the one-, two-, three-parameter logistic response model and new procedures that are suggested by resistant/robust theory are compared. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED222525

Foundations For the Mathematical Notion of Information in Item Response Theory and Robust Ability Estimation

Douglas H. Jones
Educational Testing Service
Princeton, New Jersey

Item response theory is an attempt to apply statistical theory and methodology to inferences about a subject's ability. The statistical theory receiving the most attention has been that based on the Fisherian concept of information and the associated principals and methods of maximum likelihood (Birnbaum, 1967). Forsaking Fisherian optimality to gain estimators with high stability, researcher's have recently attempted to extend item response theory to realistic testing situations by employing concepts of robust statistical theory (Wainer and Wright, 1980; Bock and Mislevy, 1981). Although robust statistical theory, as developed by Huber (1981), is primarily concerned with discovering and protecting oneself against the least informative model close by the ideal model that can explain the mechanics of inconsistent observations, item response theory researchers have developed estimators of ability that imitate the way the maximum likelihood estimator, associated with Huber's least favorable models, reduces the impact of maverick observations. While Huber's theory yields variance estimates as an easy by-product of his methods, adequate variance formulas are not available for inferences in item response theory based on robust estimators developed thus far. Extending Huber's elegant concept of robust theory to the item response problem is a worthwhile task for it may afford practitioners with streamlined techniques for making accurate inferences. In the meantime, however, our task is that of meeting the immediate problem of practitioners required to make inferences under adverse conditions, even if this is accomplished with ad hoc estimators.

This article describes an attempt to develop the concept of influence in item response theory that parallels a similar concept in robust theory in order that the methods of robust theory be better applied although the results may be ad hoc. Some of the progress in this direction has resulted in a new class (h-estimators) of robust estimators of ability and a systematic approach to the derivation of bias and variance formulas. Not only may these formulas be used for constructing confidence intervals for a subject's ability, they may be used to compare the efficiency of the h-estimators to that of the standard maximum likelihood estimator. An example will be given showing that the worse the h-estimators can do is lose 10% in efficiency, or one item in ten, relative to the maximum likelihood estimator, while maintaining strong stability in the presence of maverick observations -- a property that is absent from the behavior of the maximum likelihood estimator.

Tm 820672

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
D. H. Jones
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

After the preliminary notation and results are presented, which hopefully are adequate for informing the reader about robust methods, attention will be again directed to Huber's robust statistical theory and its applicability to item response theory will be discussed. It should be noted that one must be very careful in assuming forms of the item probability of correct response since not every model can adequately describe observed data (Holland,1981).

This article will conclude with a brief discussion on the theory of experimental design and how it provides insightful motivation for some methodology in item response theory, including the important computerized adaptive test.

Preliminaries

If $p(x;\theta)$ is the probability frequency function of the response, $X=x$, to a given item by a subject with ability θ , the Fisher information in the response X about the parameter θ is defined as $I_X(\theta) = E\{[\partial_p (X;\theta)/\partial\theta]\}^2$. The term information is used for this quantity, since the joint information in two independent responses X and Y, $I_{(X,Y)}(\theta)$ is the sum of the individual information values:

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta) , \qquad (1)$$

and the average information per observation is the last quantity divided by two. In general, the average information per observation in n item responses $X_1,\ldots, X_n$ is

$$I_n(\theta) = \Sigma I_{X_i}(\theta)/n \qquad (2)$$

where the sum is from 1 to n. It is easy to show that if only one observation is taken with equal probability from $X_1,\ldots, X_N$

where N is different from n, then $I_N(\theta)$ is the information associated with that observation. This concept also generalizes to choosing observation X_j with probability g_j to give the formula for the information per observation equal to

$$I_g(\theta) = \Sigma\, g_j I_{X_j}(\theta) . \qquad (3)$$

where the sum is from 1 to N.

It is easily observed that the information in n independent observations chosen in this fashion is $nI_g(\theta)$.

In practice, a response is either right (X=1) or wrong (X=0) and the information formula reduces to a simple form:

$$E[\partial \log p(X;\theta)/\partial\theta]^2 = (P')^2/[PQ] \qquad (4)$$

where $P \equiv P(\theta) = p(1;\theta)$, $Q = 1-P$ and $P' = \partial p(1;\theta)/\partial\theta$.

For illustration and examples, attention will be directed to the two-parameter logistic response function:

$$P(\theta) = \exp[a(\theta-b)]/[1 + \exp[a(\theta-b)] \qquad (5)$$

where $a > 0$, the discrimination parameter; $-\infty < b < \infty$, the difficulty parameter. The following results are applicable to the three-parameter models; however, since the formulas have simpler forms, the two-parameter model is used throughout.

Since $P'(\theta) = aP(\theta)Q(\theta)$, the Fisher information simplifies. Using this simplification and denoting by $P_i(\theta)$ the probability that $X_i=1$, the information in n responses is

$$I_n(\theta) = \Sigma\, a_i^2\, P_i(\theta)\, Q_i(\theta) \qquad (6)$$

and the information in n responses, chosen independently from a pool of N items such that X_j has probability g_j of being chosen, is equal to

$$nI_g(\theta) = n\, \Sigma\, g_j a_j^2\, P_j(\theta)Q_j(\theta) .$$

The log-likelihood function for n observations is equal to

$$\ell(\theta) = \Sigma[x_i \log P_i(\theta) + (1-x_i) \log Q_i(\theta)] \qquad (7)$$

and the maximum likelihood estimator $\hat{\theta}$ is the value of θ that maximizes $\ell(\theta)$. The maximum likelihood estimator is found by solving implicitly in the equation $\ell'(\hat{\theta}) = 0$; that is

$$\ell'(\hat{\theta}) = \Sigma\, a_i[x_i - P_i(\hat{\theta})] = 0. \qquad (8)$$

Example: With n = 10, b_i = - 0.8, 1.0 by .2, $a_i \equiv 1$, i = 1, 10

x = (1,1,1,1,1,0,0,0,0,0,), recursive solution of (8) yields the value $\hat{\theta}$ =0.11.

If the number of observations is sufficiently large,

$$[\hat{\theta} - \theta]/[-\ell''(\hat{\theta})]^{-\frac{1}{2}} \qquad (9)$$

has approximately the standard normal distribution thus affording an opportunity for one to construct confidence intervals that are likely to contain the true value θ. A 95% confidence interval has the form

$$\hat{\theta} - 2\,[-\ell''(\hat{\theta})]^{-\frac{1}{2}} \quad , \quad \hat{\theta} + 2\,[-\ell''(\hat{\theta})]^{-\frac{1}{2}} .$$

The variance can be shown to satisfy

$$1/-\ell''(\hat{\theta}) = 1/\Sigma\, a_i P_i'(\hat{\theta}) = 1/\Sigma a_i^2 P_i(\hat{\theta})Q(\hat{\theta}) . \qquad (10)$$

Comparing this last value with the formula for Fisher's information we see that the variance of the maximum likelihood estimator is the reciprocal of the Fisher information and this is true in general (Chernoff, 1975). In (9) the denominator may be replaced by $[nI_g(\hat{\theta})]^{-\frac{1}{2}}$ and the normal approximation continues to hold.

The important property of maximum likelihood estimation is its claim to optimality. For if T_n is any other estimator of θ that is unbiased and $V_n(\theta) = E(T_n - \theta)^2$, the mean squared error, then by the Cramer-Rao inequality, $V_n(\theta) \geq 1/nI_g(\theta)$. And therefore one can not do better in estimating than the reciprocal of the Fisher information.

Maverick Observations (Outliers)

Suppose several responses are inconsistent with the majority of the responses. For example, a subject, whose majority of responses indicate mediocre ability, gets a very difficult item correct. Thus, the maximum likelihood estimator is typically overly

influenced by the inconsistent response as the following example illustrates.

Example: Let the parameters have the same values as the previous example but let $x = (1,1,1,1,1,0,0,0,0,1)$, then $\hat{\theta} = .58$.

The primary reason for this distortion is that the implicit equation (8) weights equivalently the contribution of each item to the sum. Whenever a response is inconsistent with the rest of the responses, the difference $x_i - P_i(\hat{\theta}_{/i})$ is large ($\hat{\theta}_{/i}$ denotes the estimate obtained by deleting the ith response), and overly influencing the value of the sum.

Let $h \geq 0$ and consider the estimator that is the solution to the equation,

$$\Sigma\, a_i [x_i - P_i(\hat{\theta})][P_i(\hat{\theta})Q_i(\hat{\theta})]^h = 0 .$$

These estimators, denoted h-estimators, are not overly influenced by maverick responses. If $x_i - P_i(\hat{\theta}_{/i})$ is large, then, $[P_i(\hat{\theta}_{/i})Q_i(\hat{\theta}_{/i})]^h$ is small, so that the product does not contribute a disproportionate value to the sum.

Example: Let the parameters be as in previous examples. In the following table is displayed values of h-estimators for selected values of h.

Responses	h				
x	0(MLE)	0.5	1	2	3
1 1 1 1 1 0 0 0 0 0	0.11	0.11+	0.14	0.13	0.12
1 1 1 1 1 0 0 0 0 1	0.58	0.41	0.22	0.20	0.19

Note that h-estimators contain the maximum likelihood estimator as a special case.

The variance of the h-estimator is given by the formula

$$A(\theta;F,h) = \left[n\Sigma\, g_j a_j \overline{PQ}_j^{\,2h+1}\right] \Big/ \left[n\Sigma g_j \overline{PQ}_j^{\,h+1}[1+h(Q_j-P_j)/\overline{PQ}_j]\right]^2$$

where $F=(P_j(\theta),\ j=1,\ldots,N;g)$ and is used to denote the dependence of the variance on the underlying probability model for response and item selection, $\overline{PQ}$ denotes the product and the explicit dependence on θ is suppressed for notational convenience. Note that $A(\theta;F,0) = 1/nI_g(\theta)$.

The term efficiency will be used to denote the ratio of variance of two estimators (or in case of biased estimators, ratio of mean squared errors). If the values of the parameters are the same as in the examples, the calculated efficiency, maximum likelihood to h-estimators, are displayed for selected values of h.

h	0.5	1.0	2.0	3.0	4.0
EFF	.99	.98	.95	.92	.90

These results are most promising for they show that the h-estimators have high efficiency even for large values of h, when h-estimators are least affected by inconsistent observations.

Asymptotic variance formulas are also available for true item response probabilities that are different from the assumed forms. Denote the true item response function for item j by $P_j^*(\theta)$. Then for $F^*=(P^*_j,\ j=1,\ \ldots,\ N,g)$, the variance is given by

$$A(\theta;F^*,h) = n\left[\Sigma g_j a_j^2\, \overline{PQ}_j{}^* \overline{PQ}_j^{2h} + \Sigma g_j a_j^2\, (P_j^*-P_j)^2 \overline{PQ}_j^{2h}\right] \Big/ \left[n\Sigma\ g_j a_j^2\, \overline{PQ}_j^{h+1}[1+h(Q_j-P_j)/\overline{PQ}_j]\right]^2 .$$

Under models other than the assumed one, both the maximum likelihood estimator and the h-estimators are asymptotically biased. Formulas for the bias are given by

$$B(\theta;F^*,h) = \Sigma g_j a_j \overline{PQ}_j^{\,h}(P_j^* - P_j) / \Sigma g_j a_j^2 \overline{PQ}^{h+1} .$$

Combining the variance and square of the bias allows one to determine the relative efficiency under F^*. The only decision that must be made is which F^*. Meaningful alternatives are difficult to formulate and this is an open area needing much more exploration. More on this will be discussed later.

Example: (Distortion of Difficulty Parameters). Suppose each P^*_j has a difficulty parameter $b_j^* = b_j - 0.1$, then in the following table are displayed the relative efficiencies for selected values of h.

h	0.5	1.0	2.0	3.0	4.0
EFF	2.24	4.83	19.15	51.46	82.62

This example illustrates how astonishingly bad the maximum likelihood estimator can be relative to the ad hoc h-estimators.

Finding Suitable Models

The problem of finding models that appropriately describe maverick responses was touched upon in the last section. Should reasonable models be discovered, then neighborhoods of these models surrounding the idealmodel could be investigated for the one yielding the least Fisher information. As protection from this least favorable member one could employ the best possible estimator available for observations generated by it. In this fashion, one would naturally be lead to a "robust" methodology and this is the general program followed in standard statistical literature (Huber, 1981). Thus one is tempted to employ some of those ideas in constructing models for item responses. However, derived models may not be suitable when compared for fit on data aggretated over many subjects.

Sequential Design

As derived earlier, the Fisher information per observation X, $I_X(\theta)$, depends on the unknown ability parameters. This presents a difficulty when one tries to select an item to maximize the information. One would attempt to use some prior information to locate θ in some interval, then zero in with repeated items to yield the highest information. A method that parallels a practice in bioassay is to initially administer items widely scattered throughout the difficulty range of the item pool in order to create the prior information. Based on an estimate of $\hat{\theta}$, then, item j would be selected if $I_{X_j}(\hat{\theta})$ were largest among the remaining items.

Several ideas discussed thus far impact on this program. $I_{X_j}(\hat{\theta})$ is the quantity to maximize over j, only if the estimator is the maximum likelihood estimator. If any other kind of estimator is used, such as h-estimators or other "robust" estimators, the variance formula must be minimized by the new item that will be administered. Therefore reasonable approximations to these variance formulas are required to do the job.

Proper comparisons between sequential design schemes and proper inferences of the subject's ability may only be made if the correct repeated sampling variances or information measures are known. The value of $I_{X_i}(\hat{\theta})$ obtained at the end of a sequentially administered test is not the Fisher information of the probability response function; nor would the similarly obtained variance formula be correct for robust estimators. The problem lies in the fact that the item responses are not independent. In fact, the probability frequency function of X_{i+1} depends on the responses $X_1, \ldots, X_i$ since the item is chosen according to these responses through the value of $\hat{\theta}$. The relevant Fisher information for the i+1 observation is the expectation over $X_1, \ldots, X_i$ of the quantity

$$E[\partial_p(X_{i+1}|X_1, \ldots, X_i)/\partial\theta]^2 .$$

Once the Fisher information of an n item sequential test is obtained it should then be compared to $nI_g(\theta)$. If the information value in the sequential test is not appreciably higher than the fixed sample test, then the additional cost of sequential item administration may not be justified on the grounds of improved accuracy alone.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability (Part 5). In F.Lord & M. Novick, Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1967.

Bock, R. D., and Mislevy, R. J. Bieweight Estimates of Latent Ability, Manuscript, 1981.

Holland, P. W. When are Item Response Models Consistent with Observed Data? Psychometrika, Vol. 46, No. 1, 1981.

Huber, P. J. Robust Statistics. New York: John Wiley, 1981.

Wainer, H., and Wright, B. D. Robust Estimation of Ability in the Rasch Model, Psychometrika, Vol. 45, No. 3, 1980.